AL-TP-1993-0005

REFINEMENT OF SCORING PROCEDURES FOR THE BASIC ATTRIBUTES TEST (BAT) BATTERY

Charles E. Lance

Organizational Research and Development, Incorporated
4 114 Mink Livsey Road
Lithonia, GA 30058

Amy M. Stewart

Department of Psychology
The University of Georgia
Athens, GA 30602

Metrica, Incorporated
8301 Broadway, Suite 215
San Antonio, TX 78209

Thomas R. Carretta

HUMAN RESOURCES DIRECTORATE
MANPOWER AND PERSONNEL RESEARCH DIVISION
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352

March 1993

Interim Technical Paper for Period September 1991 - September 1992

AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS

ARMSTRONG LABORATORY

NOTICES

This technical paper is published as received and has not been edited by the technical editing staff of the Armstrong Laboratory.

Publication of this paper does not constitute approval or disapproval of the ideas or findings. It is published in the interest of scientific and technical information (STINFO) exchange.

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

Thomas R. Carretta

THOMAS R. CARRETTA
Contract Monitor

WILLIAM E. ALLEY
Technical Director
Manpower and Personnel Research Division

ROGER W. ALFORD, Lt Colonel, USAF
Chief, Manpower and Personnel Research Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1993	Interim September 1991 – September 1992

4. TITLE AND SUBTITLE

Refinement of Scoring Procedures for the Basic Attributes Test (BAT) Battery

5. FUNDING NUMBERS

C - F41689-88-D-0251
PE - 62205F
PR - 7719
TA - 18
WU - 54

6. AUTHOR(S)

Charles E. Lance
Thomas R. Carretta
Amy M. Stewart

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Organizational Research and Development, Incorporated
4114 Mink Livsey Road
Lithonia, GA 30058

Department of Psychology
The University of Georgia
Athens, GA 30602

Metrica, Incorporated
8301 Broadway,
Suite 215
San Antonio, TX 78209

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Armstrong Laboratory
Human Resources Directorate
Manpower and Personnel Research Division
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AL-TP-1993-0005

11. SUPPLEMENTARY NOTES

Armstrong Laboratory Project Scientist: Thomas R. Carretta, (210) 536-3942

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The Basic Attributes Test (BAT) is a multiple aptitude computer-based battery designed to measure individual differences in psychomotor coordination, cognitive abilities, personality, and attitudes. The Air Force plans to operationally implement the BAT as a pilot candidate selection instrument in the near future. Scores from the Air Force Officer Qualifying Test, BAT, and biographical information will be combined in a new pilot candidate selection composite to predict undergraduate pilot training outcomes. Although much useful research has been done in the BAT battery, the need for additional psychometric research to improve test scoring procedures and predictive efficiency was identified. The purpose of this study was to investigate (a) the internal consistency of item-level test scores, (b) the effects of alternative scoring procedures (e.g., treatment of outliers, data transformations, alternate scoring algorithms) on internal consistency and validity, and (c) the factor structure of the BAT. Results showed that (a) internal consistencies of most BAT scores are acceptable, indicating that the constructs are being measured reliably, (b) neither censoring outlying data points nor transforming data had a significant impact on internal consistency or validity of BAT scores, (c) few alternative scoring procedures improved BAT score validity, (d) test scores relate to a meaningful factor structure, and (e) BAT scores can be combined into an efficient model for the prediction of undergraduate pilot training performance.

14. SUBJECT TERMS

Computer-based tests
Outliers
Pilot selection
Reliability
Validity

15. NUMBER OF PAGES
46

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

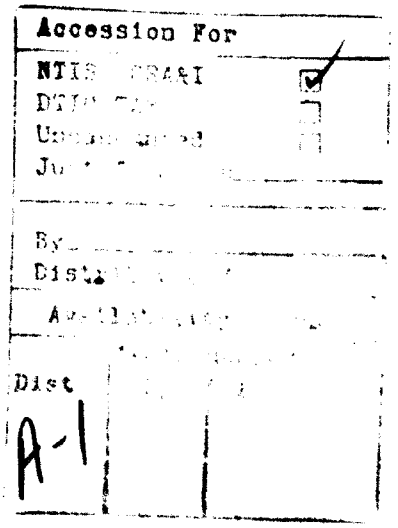
TABLE OF CONTENTS

LIST OF TABLES

PREFACE

This project was conducted under Subcontract Agreement No. 200-91-137 issued by Metrica, Inc. to Organizational Research & Development, Inc., under prime contract F41689-88-D-0251 (Delivery Order QT27, Task 57), SBA 8(a) Contract No. 68822004, issued by the Department of the Air Force.

Appreciation is extended to Mr. Cal Fresne and Sgt. Keith Weekley for their assistance in data processing, and to Drs. William E. Alley, Malcolm James Ree, and Joseph L. Weeks for their comments and technical support during this project.



REFINEMENT OF SCORING PROCEDURES FOR THE BASIC ATTRIBUTES TEST (BAT) BATTERY

SUMMARY

The Basic Attribute Test (BAT) is a computerized test battery designed to assess individual differences in psychomotor ability, cognitive abilities, personality, and activity preferences for the purpose of selection and assignment of aircrew members (Carretta, 1989, 1990a; Carretta & Siem, 1988; Kantor & Bordelon, 1985; Kantor & Carretta, 1988). Although much useful research already has been conducted on the BAT, the need for additional psychometric research to improve the battery's scoring and predictive efficiency was identified. This research addressed these needs by investigating: (a) the internal consistencies of item-level BAT scores, (b) alternative test scoring procedures, including methods for the treatment of outlying data points, data transformations, and alternative methods for forming summary scores from item-level data, and, (c) the underlying factor structure and differential validity of BAT summary scores in the prediction of Undergraduate Pilot Training (UPT) final outcome (graduation vs. elimination). Results showed that (a) internal consistencies of most BAT scores are high, indicating the homogeneity of BAT item content, and that the constructs assessed are measured reliably, (b) neither censoring outlying data points nor transforming data had a significant impact on either the internal consistency or validity of BAT summary scores, (c) few alternative scoring procedures improved the validity of BAT scores, (d) the BAT battery scores relate meaningfully to a lower-order underlying dimensionality, and (e) BAT summary scores can be combined into an efficient model for the prediction of UPT final outcome.

INTRODUCTION

Background

Measures of perceptual and psychomotor abilities were used by the U.S. Air Force (USAF) to predict flying training performance and classify aircrew members into job specialties as early as 1942 but their use was discontinued in 1955 (see Passey & McLaurin, 1966). However, in the late 1960s improvements in computer-based technology stimulated renewed interest in the development of perceptual and psychomotor tests for pilot candidate screening (Sanders, Valentine, & McGrevy, 1971). Later, in the mid 1970s, the Air Force Human Resources Laboratory (AFHRL) initiated two large-scale R&D efforts to develop and test additional computer-based pilot screening tests (Bordelon & Kantor, 1986). USAF research and development (R&D) has continued through the 1980s and into the 1990s on computerized testing of perceptual and psychomotor skills (see Bordelon & Kantor, 1986; Carretta, 1989, 1990a, 1990b, 1991; Kantor & Bordelon, 1985; Kantor & Carretta, 1988) and has recently extended into personality and attitudinal predictors of pilot training success (Carretta & Siem, 1988; Siem, 1990; Siem, Carretta, & Mercatante, 1988). Collectively, the system of computerized tests presently under R&D for the purpose of

pilot screening and classification is referred to as the Basic Attributes Test (BAT) (Carretta, 1987).

The BAT was designed to assess individual differences in psychomotor ability, cognitive abilities, personality, and activity preferences for the purpose of selection and assignment of aircrew members (Carretta, 1989, 1990a; Carretta & Siem, 1988; Kantor & Bordelon, 1985; Kantor & Carretta, 1988). Although continually evolving, the present BAT battery consists of 8 tests intended for pilot candidate screening:

(1) Two-Hand Coordination (multilimb coordination)
(2) Complex Coordination (control precision, multilimb coordination)
(3) Encoding Speed (verbal processing)
(4) Mental Rotation (spatial transformation)
(5) Item Recognition (short-term memory)
(6) Time Sharing (reaction time, rate control)
(7) Self-Crediting Word Knowledge (vocabulary and self-confidence)
(8) Activities Interest Inventory (attitudes toward risk).

Five additional tests are also administered along with the BAT selection battery and are undergoing R&D for possible classification of aircrew members:

(1) Aircrew Personality Profiler (personality test)
(2) ABCD Working Memory (verbal working memory)
(3) Anticipation (dynamic spatial ability and temporal processing)
(4) Pattern Recognition (ability to recognize visual patterns)
(5) Scanning and Allocating (compensatory tracking involving multiple tasks)

Performance on these computer-based tests is measured along several dimensions including (a) response latency, (b) response accuracy, (c) tracking error, (d) self-confidence, and (e) several personality dimensions (see Carretta, 1989; Carretta & Siem, 1988; Kantor & Bordelon, 1985; Siem, 1990, for overviews).

To date, research on the BAT battery has (a) found that Air Force Reserve Officer Training Corps (AFROTC) and Officer Training School (OTS) students achieve similar test scores (Carretta, 1990a), (b) examined the test-retest reliability of the eight selection tests (Carretta, 1991), and (c) determined that several dimensions (factors) underlie BAT scores (Bordelon & Kantor, 1986; Carretta, 1990a, 1990b). Some research also has shown that BAT summary scores have incremental validity in predicting pilot training performance when used in combination with current pilot selection instruments (e.g., Carretta, 1989), although these findings are not unequivocal (e.g., Carretta & Siem, 1988; Siem, Carretta, & Mercatante, 1988). Currently, the BAT database includes over 1100 USAF pilot candidates with BAT test scores and Undergraduate Pilot Training (UPT) final outcome (pass/fail) data.

Study Purpose

Although much research already has been conducted on the BAT database, psychometric work which could be potentially useful for operational use of BAT scores in terms of increasing its scoring and predictive efficiency has yet to be completed. The purposes of the present study were to investigate:

(a) Internal consistencies of item-level BAT scores, with implications for shortening or lengthening tests;

(b) Alternative test scoring procedures, including methods for the treatment of outlying data points, data transformations, and alternative methods for forming summary scores from item-level data, with implications for increased scoring efficiency of the BAT; and,

(c) The underlying factor structure and differential validity of BAT summary scores, in order to identify efficient predictors of pilot training success (UPT final outcome).

METHOD

Literature Review

Initially, relevant literature was reviewed, including literature relating to (a) development and evaluation of the BAT (e.g., Carretta, 1989, 1990b; Kantor & Carretta, 1988; Siem et al., 1988), (b) psychometrics and test development procedures (e.g., Crocker & Algina, 1986; Lord & Novick, 1968; Wainer & Braun, 1988), (c) alternative test scoring procedures (e.g., Crocker & Algina, 1986; Greaud & Green, 1986), (d) computerized testing (Cory, Rimland, & Bryson, 1977; Green, 1988), (e) personality predictors of pilot performance (e.g., Dolgin & Gibb, 1989; Novello & Youssef, 1974), (f) development and evaluation of regression and prediction models (e.g., Cohen & Cohen, 1983; Skinner, 1978), (g) reliability theory (e.g., Carmines & Zeller, 1979; Feldt & Brennan, 1989; Lord & Novick, 1968), (h) the treatment of outliers and influential data points (e.g., Chatterjee & Hadi, 1986), (i) data transformations for nonnormal data distributions (e.g., Stevens, 1986), (j) cognitive information processing models (e.g.,Ackerman, 1986; Adams, 1987), (k) speed/accuracy tradeoffs in information processing (e.g., Link, 1982; Pachella, 1974), and other related literature (e.g., Campbell, 1991; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990). Theoretical developments and practical recommendations in these literatures guided the analyses and interpretations described in subsequent sections.

Database

Data for the present study included item-level data on 11 of the 13 tests. Self-Crediting Word Knowledge test data were not included for analyses because summary scores had been found previously to penalize good performance in regression models developed to predict UPT final outcome. Data on the Aircrew

Personality Profiler also were not included because they were the subject of a separate R&D effort. Thus, the present study included data on the remaining 11 BAT tests, UPT final outcome measures, and additional predictors of pilot training success (Air Force Office Qualifying Test [AFOQT] scores and amount of previous flying experience). All data were obtained from AFROTC and OTS pilot trainees who had been chosen for UPT, in part, based on their AFOQT scores. All subjects had completed a 4-year college degree prior to entering UPT and ranged in age between 21 and 27 years. Subjects were informed that their performance on the BAT battery would neither affect their continuation in UPT nor be entered into their permanent service records. Details of the BAT instrumentation and data collection procedures for the database examined in the present study are given in Carretta (1987, 1989, 1990a, 1990b) and Kantor and Carretta (1988).

Sample sizes for the 11 BAT tests studied here are shown in Table 1, which also lists the performance dimensions which are scored for each test. Sample sizes vary because different combinations of tests were administered at different times and data collection sites, and due to differing degrees of missing data (see below).

Data were coded missing if (a) original, study-generated missing values were encountered (i.e., "9," "0," or "blank"), (b) data indicated that the subject was inactive during a portion of the test (e.g., control stick movement rates indicated no movement), or (c) response latencies were less than 200 milliseconds (this is an asymptotic lower-bound for simple reaction times, see Luce, 1986). Except as indicated in the following sections, data were deleted elementwise (i.e., data records were not deleted on a listwise basis).

ANALYSES AND RESULTS

Internal Consistencies of Item-Level BAT Scores

As is shown in Table 1, multiple performance dimensions were scored for most of the tests (e.g., latency and accuracy of performance). Furthermore, most of the tests consist of a large number of items, some of which might not be homogeneous with respect to the underlying construct being measured. Thus, it was sought to determine, at the item-level of analysis, the internal consistency of test items.

Previous research on the internal consistency of the BAT battery has reported Cronbach alphas in the 0.90s for response latency and tracking error performance measures, but the internal consistencies of accuracy measures have been found to be lower (Carretta, 1991). The latter finding likely reflects the effects of attenuation due to ceiling effects typically encountered on highly speeded tests, that is, on tests containing easy items which are scored for the number of items answered or, relatedly, speed of response.

TABLE 1. SAMPLE SIZES FOR BAT TESTS

Acronym	Test Name and Performance Dimensions	Sample Size
BAT Selection Battery:		
PS2	Two-Hand Coordination	2451
	PS2XA - Horizontal Tracking Error PS2YA - Vertical Tracking Error	
PS2	Complex Coordination	2451
	PS2XB - Horizontal Tracking Error PS2YB - Vertical Tracking Error PS2ZB - Rudder Bar Tracking Error	
ENC	Encoding Speed	2470
	ENCRT - Response Latency ENCRO - Response Outcome	
MRT	Mental Rotation	2147
	MRTRT - Response Latency MRTRO - Response Outcome	
ITM	Item Recognition	2209
	ITMRT - Response Latency ITMRO - Response Outcome	
TMS	Time Sharing	2536
	TMSRT - Response Latency TMSLD - Tracking Difficulty	
AII	Activities Interest Inventory	2353
	AIIRT - Response Latency AIIRO - Response Outcome	
Experimental Tests:		
ABC	ABCD Working Memory	377
	ABCST - Study Time ABCRT - Response Latency ABCRO - Response Outcome ABCCF - Confidence Rating	

TABLE 1. SAMPLE SIZES FOR BAT TESTS (CONT'D)

Acronym	Test Name and Performance Dimensions	Sample Size
ANT - Anticipation		967
	ANTTE - Target Error	
PAT - Pattern Recognition		934
	PATRT - Response Latency PATRO - Response Outcome	
SAA - Scanning and Allocating		946
	SAASW - Quadrant Switches SAAER - Tracking Error	

A number of different statistical procedures were reviewed for the estimation of tests' internal consistencies (Bentler & Woodward, 1983; Feldt & Brennan, 1989; Hattie, 1985, Wittman, 1988). Generally, Cronbach's (1951) alpha coefficient or one of Kuder and Richardson's (1937) formulae is recommended to assess test internal consistency, although some (e.g., Crocker & Algina, 1986) have recommended split half correlations for speeded tests. However, most internal consistency coefficients are lower bounds to, and special cases of Cronbach's (1951) alpha coefficient (see Feldt & Brennan, 1989) which is theoretically the mean of all possible split-half reliability estimates. For these reasons, and because Cronbach's alpha is an appropriate index of test item homogeneity, it was used here to estimate the internal consistency of response outcome (accuracy), tracking error, response latency, self-confidence, and personality scores.

Coefficients alpha for test items are shown in Table 2 separately for the different performance dimensions scored. Alpha coefficients generally were very high (e.g., see Nunnally, 1970). With the exception of several of the response outcome measures (right vs. wrong) and average response time from Time Sharing (TMSRT), all coefficients were .90 or higher. Coefficient alpha should not be interpreted as an index of unidimensionality (Hattie, 1985), but the coefficients alpha in Table 2 do indicate that BAT tests generally are homogeneous with respect to item content.

TABLE 2. COEFFICIENTS ALPHA FOR BAT SELECTION BATTERY AND EXPERIMENTAL TESTS

Test	Dimension	No. of Items	Alpha
1. Two-Hand Coordination	PS2XA	10	.99
	PSZYA		.99
2. Complex Coordination	PS2XB	10	.96
	PS2YB		.96
	PS2ZB		.95
3. Encoding Speed	ENCRO	96	.70
	ENCRT		.96
4. Mental Rotation	MRTRO	72	.89
	MRTRT		.97
5. Item Recognition	ITMRO	48	.52
	ITMRT		.96
6. Time Sharing	TMSRT	35	.88
	TMSLD	190	.99
7. Activities Interest Inventory	AIIRO	81	.87
	AIIRT		.95
8. ABCD Working Memory	ABCRO	48	.94
	ABCST		.97
	ABCRT		.92
	ABCCF		.97
9. Anticipation	ANTTE	50	.92
10. Pattern Recognition	PATRO	30	.61
	PATRT		.92
11. Scanning & Allocating	SAASW	180	.90
	SAAER	450	.98

Note. For scores involving tracking performance (Two-Hand Coordination, Complex Coordination, Time Sharing, and Scanning and Allocating) "No. of Items" refers to number of time intervals. See Table 1 for definitions of Test/Dimension acronyms.

Recommendations for Shortening or Lengthening Tests.

Based on results in Table 2, estimates of the internal consistencies of shorter or longer versions of the tests can be derived on the basis of the Spearman-Brown prophesy formulae (see Lord & Novick, 1968). Also, for a test with N items and reliability $r_{xx'}$, the number of test items K required to achieve a desired reliability $R_{xx'}$ can be estimated from:

$$K = N^{*}\frac{R_{xx'}(1 - r_{xx'})}{r_{xx'}(1 - R_{xx'})}. \qquad (1)$$

K was estimated for BAT performance dimensions for $R_{xx'} = .70$, considered by some to be adequate for exploratory research (Nunnally, 1970), through $R_{xx'} = .95$, in increments of .05. Results are shown in Table 3.

Disregarding the response outcome performance dimensions (for which no suitable reliability estimates were identified), and Time Sharing tracking difficulty (TMSLD), where "items" are not experimentally independent, Spearman-Brown estimates suggest that if the desired $R_{xx'} = .90$, then all but the Time Sharing test could be shortened, by as much as 72% (Mental Rotation), and on the average by approximately 50%. On the other hand, these results also can be interpreted as indicating that the tests, which already are relatively brief, achieve a high degree of internal consistency which need not be compromised by additional shortening. Practically, the psychometric results in Tables 2 and 3 should be interpreted in the context of operational (e.g., time and budgetary) constraints.

In summary, results on the internal consistency of the tests indicated that (a) the internal consistencies of response outcome performance dimensions were low, pointing to the lack of a suitable index of internal consistency for these dimensions, (b) internal consistencies of the remaining measures were nearly uniformly high, suggesting that the test items are homogeneous with respect to their item content, and (c) most tests could be shortened, and some considerably, to achieve a desired internal consistency of $R_{xx'} = .90$.

TABLE 3. SHORTENING VS. LENGTHENING BAT TESTS

Test/ Dimension	Orig. No. of Items	Desired $R_{xx'}$ .70	.75	.80	.85	.90	.95
AIIRO	81		33	44	63	100	210
AIIRT	81			17	24	38	81
ENCRO	96	96	123	165	233	370	782
ENCRT	96		12	16	23	36	76
ITMRO	48		20	26	37	59	124
ITMRT	48				11	18	38
MRTRO	72		14	39	56	88	187
MRTRT	72			9	11	20	42
PS2XA	10			1	2	3	6
PSZYA	10			1	2	3	6
PS2XB	10			1	2	4	8
PS2YB	10			1	2	4	8
PS2ZB	10		1	2	3	5	10
TMSRT	35		14	19	27	43	91
TMSLD	190				11	17	36
ABCRO	48			12	17	28	58
ABCST	48				8	13	28
ABCRT	48			17	24	38	79
ABCCF	48				8	13	28
ANTTE	50			17	25	39	83
PATRO	30	45	58	77	109	173	364
PATRT	30			10	15	23	54
SAASW	180		41	54	77	122	257
SAAER	450			37	52	83	174

Note. For scores involving tracking performance (Two-Hand Coordination, Complex Coordination, Time Sharing, and Scanning and Allocating) "No. of Items" refers to number of time intervals. See Table 1 for definitions of Test/Dimension acronyms.

Evaluation of Alternative Test Scoring Procedures

To date, a limited number of scoring systems have been applied to BAT performance measures. Most of these have involved the computation of summary scores based on measures that are routinely recorded (e.g., percent correct, mean response latency, cumulative tracking error, see Carretta, 1991). Research on the BAT summary scores has shown that some of them (a) are significantly positively skewed and platykurtic, (b) contain outlying data points, and (c) fail to contribute incrementally

in predicting UPT final outcome (Carretta, 1987, 1989; Kantor & Carretta, 1988). The research described in this section was intended to address these issues by investigating the effects of (a) data transformations and procedures for the treatment of outlying data points, and (b) alternative procedures for forming summary scores on the basis of item-level BAT data, on the internal consistency of summary scores and on the predictability of UPT final outcome. Specifically, the effects of alternative test scoring procedures were examined in terms of:

(a) maximizing the predictability of UPT final outcome (graduation vs. elimination);

(b) reducing the number of predictor scores in an overall model predictive of UPT outcome (e.g., by combining different performance measures into a single score indicating overall performance);

(c) insuring that test scoring procedures are consistent with test instructions (e.g., to reflect speed-accuracy tradeoffs in subjects' performance strategies).

Data Transformations and Outlier Deletion.

Response latency data typically are positively skewed (Wickens, 1984) and this is true of BAT response latency and tracking error measures (Carretta, 1987, 1989). As is described by Luce (1986) response latency data are often routinely subjected to some form of data transformation to approximately normalize the data for statistical procedures which are based on normal theory.

Stevens (1986) has reviewed the effects of various transformations on nonnormal data (e.g., multimodal, skewed, and/or kurtic data). Although there is some debate on the appropriateness of effecting data transformations (e.g., see Games, 1983, 1984; Levine & Dunlap, 1983), Stevens (1986) and Mosteller and Tukey (1977) have demonstrated the effectiveness of several transformations in near-normalizing nonnormal data.

Transformations that appear to be most effective for positively skewed and platykurtic data distributions, and those that were applied here to the BAT data, are the square-root and natural logarithm transformations (Stevens, 1986). The effects of these transformations, applied both at the item-level scores as well as the summary score-level of analysis, were evaluated in terms of their influence on the internal consistency of BAT scores and on the relationships between BAT scores and the UPT final outcome.

Related to the problem of normality (Pedhazur, 1982) is the problem of outliers and influential data points (OIDPs) (Belsley, Kuh, & Welsh, 1980). There are a number of procedures for detecting OIDPs (see Belsley et al., 1980; Chatterjee & Hadi, 1986; Chatterjee & Price, 1977; and Stevens, 1984, 1986 for reviews). Some of these are statistically based, while others are more ad hoc. Previous approaches to the treatment of outliers in BAT data include the deletion of data lying outside six standard deviations from the mean (Bordelon & Kantor, 1986), and recoding observations that

lay outside three standard deviations from the mean to be equal to exactly three standard deviations from the mean (Carretta, 1990a).

According to Chatterjee and Hadi (1986), a "bewilderingly large number of statistical quantities have been proposed to study outliers" (p. 379). Based on their review, they concluded that "only three of these measures along with some graphical displays...provide...a complete picture of outliers" (p. 379). In particular, they found that examination of univariate data plots, though simple, was an effective means for identifying OIDPs, and this was the approach taken here. Specifically, univariate frequency distributions were examined for discontinuities in both item-level and summary score-level BAT scores to identify OIDPs for deletion. In general, decision rules for the deletion of OIDPs corresponded approximately to the 1st (negatively skewed data) or 99th percentile (positively skewed data). Specific decision rules are presented in Appendix A. To summarize, the effects of both data transformations and the deletion of OIDPs, both at the item and summary score levels, were assessed on BAT scores' internal consistencies and correlations with UPT final outcome.

Results for BAT summary scores' internal consistencies are shown in Table 4. The first column in Table 4 shows coefficients alpha for the original, untransformed and uncensored data. The next two columns show alpha coefficients with a square-root (SQRT) or natural logarithm (LOG) transformation applied to item-level data. The fourth column shows effects of deleting outliers at the item-level, and the last two columns show the combined effects on BAT scores' internal consistencies of transforming data and deleting outliers at the item-level of analysis. Finally, mean coefficients alpha are shown in the bottom row of Table 4. Results in Table 4 show little beneficial effects either of data transformations or outlier deletion which actually lowered coefficients alpha in many cases.

Results for BAT summary scores' correlations with UPT final outcome are shown in Table 5. The first column of Table 5 shows correlations between UPT final outcome and the original, untransformed and uncensored data. The next four columns show similar results for BAT data with a square root (SQRT) or logarithm (LOG) transformation applied at the item- (r_{ITM}) and the total (summary) score level (r_{TOT}). The next two columns show results with outliers deleted at the item level (r_{ITM}) and the total score level (r_{TOT}), and the last four columns show results for data in which both outliers were deleted and data transformations were effected. Finally, mean absolute values of the correlations in Table 5 are shown in the bottom row. As Table 5 shows, no one strategy for transforming data and/or deleting outliers could be recommended for increasing the predictability of UPT final outcome for all BAT summary scores. Thus, results were examined separately for each BAT summary score to explore optimum scoring strategies.

TABLE 4. COEFFICIENTS ALPHA FOR ORIGINAL, OUTLIER-DELETED, AND TRANSFORMED DATA

Test/ Dimension	Original Data	Transformed Data		Outliers Deleted	Outliers Deleted & Transformed	
		SQRT	LOG		SQRT	LOG
AIIRO	.87	.87	.87	.88	.88	.88
AIIRT	.95	.97	.97	.96	.96	.97
ENCRO	.70	.70	.70	.67	.68	.68
ENCRT	.96	.97	.98	.96	.97	.97
ITMRO	.77	.77	.69	.69	.65	.65
ITMRT	.98	.99	.98	.98	.98	.99
MRTRO	.95	.95	.95	.94	.97	.94
MRTRT	.98	.99	.99	.98	.94	.99
PS2XA	.99	.99	.99	.99	.98	.98
PSZYA	.99	.99	.99	.98	.99	.99
PS2XB	.96	.97	.97	.95	.96	.97
PS2YB	.96	.97	.97	.94	.96	.97
PS2ZB	.95	.96	.96	.93	.95	.95
TMSRT	.88	.90	.89	.86	.88	.89
TMSLD	.99	.99	.99	.99	.98	.99
ABCRO	.94	.94	.94	.89	.89	.89
ABCST	.97	.98	.99	.95	.97	.97
ABCRT	.92	.96	.97	.89	.89	.90
ABCCF	.97	.97	.97	.96	.96	.96
ANTTE	.92	.93	.94	.85	.84	.82
PATRO	.61	.61	.61	.50	.50	.50
PATRT	.92	.93	.95	.88	.89	.89
SAASW	.93	.94	.94	.93	.94	.94
SAAER	.98	.85	.83	.98	.85	.83
MEAN	.92	.92	.92	.90	.89	.90

TABLE 5. DIMENSION SCORE CORRELATIONS WITH UPT FINAL OUTCOME

Test/ Dimension	Original Data	Transformed Data				Outliers Deleted		Transformed Data & Outliers Deleted			
		SQRT		LOG				SQRT		LOG	
	r	r_{ITM}	r_{ITOT}	r_{ITM}	r_{ITOT}	r_{ITM}	r_{ITOT}	r_{ITM}	r_{ITOT}	r_{ITM}	r_{ITOT}
AIIRO	.01	.01	.01	.01	.02	.01	.01	.01	.01	.01	.01
AIIRT	-.02	.00	-.02	.01	-.03	-.03	-.02	.00	-.02	.00	-.03
ENCRO	.05**	.05**	.05**	.05**	.05**	.05**	.06**	.05**	.06**	.05**	.06**
ENCRT	-.03	-.07**	-.03	-.01	-.03	-.03	-.03	-.03	-.03	-.00	-.03
ITMRO	.03	.03	.03	.02	.03	.03	.04**	.03	.04*	.03	.04
ITMRT	-.11**	-.00	-.11**	.01	-.11**	-.11**	-.10**	.00	-.10**	.00	-.10**
MRTRO	.04	.03	.03	.03	.03	.04	.05**	.02	.04	-.02	.04
MRTRT	-.07**	-.07**	-.08**	-.03	-.07**	-.07**	-.07**	-.04	-.06**	-.02	-.06**
PS2XA	-.18**	-.20**	-.20**	-.21**	-.21**	-.20**	-.19**	-.20**	-.21**	-.22**	-.21**
PSZYA	-.19**	-.20**	-.20**	-.20**	-.19**	-.18**	-.17**	-.18**	-.17**	-.18**	-.17**
PS2XB	-.08**	-.08**	-.09**	-.07**	-.08**	-.08**	-.09**	-.07**	-.09**	-.06**	-.08**
PS2YB	-.09**	-.10**	-.10**	-.11**	-.11**	-.10**	-.12**	-.10**	-.12**	-.10**	-.12**
PS2ZB	-.11**	-.12**	-.12**	-.11**	-.12**	-.11**	-.12**	-.11**	-.12**	-.10**	-.12**
TMSRT	-.21**	-.23**	-.22**	-.24**	-.22**	-.20**	-.22**	-.21**	-.22**	-.20**	-.22**
TMSLD	.08**	.08**	.08**	.08**	.08**	.08**	.09**	.08**	.09**	.05*	.09**
ABCRO	.04	.04	.04	.04	.03	.04	.07	.04	.07	.04	.07
ABCST	-.10*	-.09*	-.09*	-.08*	-.08*	-.10*	-.07	-.10*	-.07	-.09*	-.07
ABCRT	-.06	-.04	-.07	.02	-.06	-.04	-.09*	.00	-.08*	.00	-.07
ABCCF	.04	.04	.04	.03	.04	.04	.01	.04	.01	.03	.00
ANTTE	-.00	-.01	-.00	-.02	-.01	-.06*	.03	.08	.03**	.09*	.02
PATRO	.06*	.06*	.05*	.06*	.04	.06*	.09**	.06*	.09**	.06*	.09**
PATRT	-.06*	-.06*	-.01	-.05	-.02	-.02	-.00	-.03	.00	-.04	-.00
SAASW	.09**	.09**	.09**	.09**	.07**	.05	.09**	.07*	.09**	.07*	.09**
SAAER	-.11**	-.10**	-.13**	-.10**	-.14**	-.15**	-.20**	-.09**	-.19**	-.09**	-.18**
MEAN	|.08|	|.08|	|.08|	|.07|	|.08|	|.07|	|.08|	|.07|	|.08|	|.06|	|.08|

* $p < .05$; ** $p < .01$

Table 6 shows coefficients alpha and summary scores' correlations with UPT final outcome for the untransformed and uncensored data along with similar results for data rescored according to the recommended procedure (results for All are discussed separately in the following section). The bottom row of Table 6 containing mean correlations shows that the beneficial effects of data censoring and transformations on validities (correlations with UPT final outcome) were modest at best, and were actually attained at the expense of a slight decrease of summary scores' internal consistencies. The strategy most often recommended was the deletion of outlying total scores, and is the option most likely to deal effectively with subjects who do not perform according to test instructions. Alternately, in an operational system, subjects who do not perform according to test instructions should be assigned a maximum valid score (i.e., a "fenced" score, see Appendix A).

TABLE 6. EFFECTS OF DATA TRANSFORMATIONS AND DATA CENSORING

Test/Dimension	Coefficient Alpha		r With UPT Final Outcome		Recommendation
	Orig.	New	Orig.	New	
AllRO	.87	-	.01	-	(See text)
AllRT	.95	-	-.02	-	(Seetext)
ENCRO	.70	.70	.05**	.05**	Sqrt-Itm
ENCRT	.96	.97	-.03	-.07**	Sqrt-Itm
ITMRO	.77	.69	.03	.04*	Out-Tot
ITMRT	.98	.98	-.11**	-.10**	Out-Tot
MRTRO	.95	.94	.04*	.05**	Out-Tot
MRTRT	.98	.98	-.07**	-.07**	Out-Tot
PS2XA	.99	.99	-.18**	-.21**	Log-Tot
PSZYA	.99	.99	-.19**	-.19**	Log-Tot
PS2XB	.96	.95	-.08**	-.09**	Out-Tot
PS2YB	.96	.94	-.09**	-.12**	Out-Tot
PS2ZB	.95	.93	-.11**	-.12**	Out-Tot
TMSRT	.88	.86	-.21**	-.22**	Out-Tot
TMSLD	.99	.99	.08**	.09**	Out-Tot
ABCRO	.94	.89	.04	.07	Out-Tot
ABCST	.97	.97	-.10*	-.10*	None
ABCRT	.92	.89	-.06	-.09*	Out-Tot
ABCCF	.97	.97	.04	.04	None
ANTTE	.92	.82	.00	.09**	Out-Itm& Log-Itm
PATRO	.61	.61	.06*	.06*	Sqrt-Itm
PATRT	.92	.93	-.06*	-.06*	Sqrt-Itm
SAASW	.93	.93	.09**	.09**	Out-Tot
SAAER	.98	.98	-.11**	-.20**	Out-Tot
MEAN	.92	.90	|.08|	|.10|	

*$p < .05$; ** $p < .01$

In summary, neither data transformations nor the deletion of outliers, either at the item- or total score-level of analysis, had dramatic effects on BAT scores' (a) internal consistencies, or (b) correlations with UPT final outcome. One alternative scoring strategy which yielded modest validity increments, and which had minimal adverse effects on summary scores' internal consistencies was the deletion of outlying scores at the summary score level.

Alternative Scoring Methods.

As was mentioned earlier, a limited number of alternative scoring systems have been applied to measures that are routinely collected from the administration of the BAT battery. In addition to the dimension scores referred to previously, Carretta (1989, 1990b) also computed additional derived scores in a subset of the tests examined here, including (a) average response latency for correct responses only, (b) performance cross-product scores (e.g., percent correct x average response latency), and (c) response latency standard deviations across trials. We examined these, plus additional scoring methods, including:

(a) "Formula scores" which correct for guessing, for example: $Xf = (R - W)/(k - 1)$, where Xf refers to the "corrected" score, R to the number of items answered correctly, W to the number wrong answers, and k to the number of response alternatives for each item (Crocker & Algina, 1986; Lord & Novick, 1968; Ree, 1976);

(b) Cross-product terms which reflect tradeoffs in subject response strategies, for example:

- Xcm = Pc*Mc, where the cross-product score Xcm weights percent correct (Pc) by a standardized response latency measure (i.e., z-score) for items answered correctly (Mc); and Xcwm = Pc*Mc - Pw*Mw where the cross-product score Xcwm also effects a correction for the percentage of incorrect answers (Pw) by their standardized response latencies (Mw);

- Xcc = Pc*Cc where the cross-product Xcc weights percent correct (Pc) by a standardized confidence rating (Cc) and Xcwc = Pc*Cc - Pw*Cw which also corrects for incorrect responses;

(c) Ollman's (1966) measure of performance which corrects for fast guessing in a choice reaction time task:

$$Xfg = (Pc*Mc - Pw*Mw)/(Pc - Pw), \quad (2)$$

where Pc and Pw refer to percent correct and incorrect responses, respectively, and Mc and Mw are the associated response latencies. According to Link (1982) and Yellott (1967, 1971), this index reflects the mean response latency for correct responses corrected for "fast guess" responses which are equally likely to be correct or incorrect. However, one practical problem with Ollman's

(1966) index is that it may encourage a tradeoff of speed for accuracy, that is, Xfg is maximized when both Pc and Mc are high.

(d) An adaptation of Ollman's (1966) index which weights accuracy by response speed (S, or reciprocal latency), rather than response latency:

$$X_{SA} = (Pc*Sc - Pw*Sw)/(Pc - Pw). \quad (3)$$

Here, X_{SA} represents the speed of correct responses corrected for fast guessing, Pc and Pw represent the percentages of correct and incorrect responses, and Sc and Sw represent mean standardized scores for the speed of correct and incorrect responses, respectively. Thus, this scoring system encourages both quick and accurate responses.

(e) Scoring only certain portions of BAT tests. This is based on the idea that performance tends to stabilize following an initial task learning or acclimatization period.

Correlations between UPT final outcome and alternative summary scores were explored to determine which, if any, held promise for (a) summarizing relations between UPT final outcome and multiple overall dimension scores, and (b) increasing the predictability of the UPT final outcome. Results are summarized in Table 7, which lists summary scores which are routinely calculated from the administration of the BAT battery, as well as alternative summary scores which best predicted UPT final outcome.

In many cases, the best predictors of UPT final outcome were the routinely calculated summary scores. Briefly, although a large number of alternative scoring procedures were applied to BAT data, few resulted in improved predictability of UPT final outcome. Specifically, correlations of only certain portions of tests (e.g., first or last block of items) were less valid predictors of UPT final outcome than were total scores. Two exceptions were All scores based on valid items identified through traditional item analytic procedures (e.g., Crocker & Algina, 1986). The majority of derived composites and cross-products also offered little improvement in validity (exceptions are noted in Table 7). Finally, fast guess model scores generally were ineffective in predicting UPT final outcome.

In summary, a large number of alternative scoring methods were examined to explore whether they (a) more efficiently summarize relations between BAT summary scores and UPT final outcome, and (b) increase the predictability of UPT final outcome, as compared to routinely computed BAT summary scores. With the exceptions noted in Table 7 (which should be cross-validated in future research), alternative scoring procedures largely were ineffective.

TABLE 7. ALTERNATIVE SCORING METHODS

Test		Summary Score	r with UPT Final Outcome
All	R -	Percent of risky alternatives chosen	.01
	R -	Response latency	-.01
	A -	Response outcome for the 10 most valid items	.10**
	A -	Response outcome for 21 items with correlations ≥ .02 with UPT final outcome	.09**
ENC	R -	Percent correct	.05**
	R -	Response latency	-.07**
	A -	Response speed x percent correct	.08**
ITM	R -	Percent correct	.03
	R -	Response latency	-.09**
	A -	Response speed x percent correct	.10**
MRT	R -	Percent correct	.06**
	R -	Response latency	-.09**
	A -	Response latency for correct responses	-.06**
	A -	Response latency x percent correct	-.06**
PS2	R -	Two-Hand Coordination X-axis error	-.21**
	R -	Two-Hand Coordination Y-axis error	-.19**
	R -	Complex Coordination X-axis error	-.09**
	R -	Complex Coordination Y-axis error	-.12**
	R -	Complex Coordination rudder error	-.12**
	A -	Sum of PS2 summary scores.	-.22**
TMS	R -	Response latency	-.20**
	R -	Difficulty level	.09**
	A -	Response speed x Difficulty level	.19**
ABC	R -	Percent correct	.07
	R -	Study time	-.10*
	R -	Response latency	-.05
	R -	Confidence rating	.04
	A -	Study time last 1/2 of test	-.10*
	A -	Response latency last 1/2 of test	-.11*
ANT	R -	Total error	.09**
PAT	R -	Percent correct	.06*
	R -	Response latency	-.05*
	A -	Fast guess score	-.06*
SAA	R -	Number of switches	.09**
	R -	Tracking error	-.11**

Note. "R" indicates a summary score routinely calculated from the administration of the BAT; "A" indicates an alternative scoring procedure.

* $p < .05$; ** $p < .01$

BAT Factor Structure and Differential Validity

Although the BAT tests are designed to assess different aspects of interests, and cognitive, perceptual, and psychomotor abilities (Carretta, 1987), some of the tests are scored along similar dimensions of performance (e.g., response latency and response accuracy). Thus, the underlying dimensionality and the differential validity of the test battery for the prediction of UPT final outcome remains to be determined. These were the purposes of the research described in this section.

Dimensionality of BAT Summary Scores

Although there is some debate over the appropriateness of various procedures for assessing the unidimensionality of test items (Hattie, 1985), factor analysis is accepted as appropriate for modeling the latent dimensionality underlying relations among multiple test battery scores (Mulaik, 1972, 1988). Factor analysis (FA) can be conducted in an exploratory or confirmatory mode, depending on whether theoretically motivated restrictions can be imposed upon estimates of model parameters (e.g., factor loadings, uniquenesses, and/or factor covariances, see Long, 1983). In practice, principal components analyses (PCAs) often are recommended in lieu of factor analyses for exploratory analyses due to less stringent statistical assumptions and computational ease. Since the theory of the structure underlying the BAT system is not strong, its underlying dimensionality was investigated using exploratory PCA.

Due to sample size requirements, data for the seven BAT selection tests and the four experimental tests were analyzed separately (e.g., listwise deletion of missing data on all tests resulted in $N = 21$). Thus PCA solutions are reported separately for the selection and experimental tests. In both solutions (a) components were extracted by the principal axes method (Mulaik, 1972), (b) the number of components retained was determined jointly on the basis of the number of eigenvalues greater than 1.0, inspection of the scree plot, and interpretability (Cattell, 1966; Zwick & Velicer, 1982, 1986), and (c) components were rotated to an oblique solution using the oblimin criterion (Mulaik, 1972). Also in order to retain the largest samples possible for analyses, PCAs were performed both on data in which (a) missing data were deleted pairwise, and (b) outlying summary score values were replaced by maximum (minimum) values specified earlier in the section on outlier treatment (i.e., outlying values were "fenced" at the maximum valid values. For example, any score greater than the maximum valid value of 9000 would be assigned the value 9000).

A five-component solution was retained for the BAT selection data. Since the solutions were nearly identical for the pairwise deleted and fenced-value data sets, only the latter is reported. Principal component (PC) loadings for the test scores are shown in Table 8. These components were interpreted as:

I. Two-Hand Coordination
II. Complex Coordination
III. Response Latency
IV. Response Accuracy
V. Risk Taking

TABLE 8. FIVE COMPONENT SOLUTION FOR BAT SELECTION TESTS

Test/ Dimension	Principal Component I	II	III	IV	V
AIIRO	.01	-.07	-.02	.04	.71**
AIIRT	.05	-.06	-.08	.08	-.79**
ENCRO	-.01	.04	.20	.75**	.05
ENCRT	.00	.03	.81**	.26	.03
ITMRO	-.12	-.04	.13	.62**	-.15
ITMRT	.19	-.01	.77**	.03	.08
MRTRO	.11	-.02	-.24	.70**	.04
MRTRT	-.09	.00	.85**	-.12	-.08
PS2XA	.96**	.06	-.05	.04	.03
PS2XB	-.06	.94**	-.03	.04	.00
PS2YA	.98**	-.01	-.05	.06	.00
PS2YB	-.01	.92**	-.08	.03	-.03
PS2ZB	-.00	.85**	.02	-.01	.04
TMSRT	.62**	-.09	.19	-.12	-.11
TMSLD	-.14	-.33*	-.23	.12	.07

*loading $\geq |.30|$; **loading $\geq |.50|$.

PC correlations are shown in Table 9. Despite the fact that PCs were rotated to an oblique solution, PCs were essentially uncorrelated, suggesting that the dimensions underlying BAT selection battery scores were nearly independent.

TABLE 9. CORRELATIONS AMONG BAT SELECTION TEST PCS

Component	I	II	III	IV	V
I	1.00				
II	.06	1.00			
III	.21	.19	1.00		
IV	-.09	-.13	.02	1.00	
V	-.11	-.08	-.14	-.00	1.00

A two-component solution was retained for the experimental test data. Again, results for the pairwise-deleted and fenced-value data sets were virtually identical so only the latter are shown. PC loadings are shown in Table 10. Component loadings did not achieve as simple a structure for the experimental tests as they did for the BAT selection battery. Nevertheless, PCs were interpretable as:

I. Working Memory
II. Resource Allocation

The correlation between these two components was .14, indicating that they were distinct dimensions underlying experimental test performance.

In summary, PC analyses resulted in interpretable solutions for both the BAT selection and experimental test data sets, suggesting that meaningful lower-order dimensionalities underlay both sets of performance dimension scores. These results also suggested additional ways in which BAT scores could be combined for the purpose of predicting UPT final outcome.

TABLE 10. TWO COMPONENT SOLUTION FOR EXPERIMENTAL TESTS

Test/ Dimension	Principal Component I	Principal Component II
ABCCF	.76**	-.20
ABCRO	.72**	-.26
ABCRT	.80**	.15
ABCST	.76**	.33*
ANTTE	.50**	-.31*
PATRO	.42*	-.25
PARRT	.54**	.15
SAASW	-.04	-.72**
SAAER	-.03	.80**

*loading $\geq |.30|$; **loading $\geq |.50|$

Differential Validity of BAT Summary Scores

One finding from earlier research on the BAT battery is that not all BAT dimension scores contribute incrementally to the prediction of UPT final outcome (Bordelon & Kantor, 1986; Carretta, 1989, 1990b), despite the fact that most summary scores' zero-order correlations with UPT final outcome (though not large) are in the expected direction and statistically significant. Results presented in the previous two sections suggested ways in which summary scores could be combined for more efficient prediction of UPT final outcome (in terms of a reduced number of predictors).

As in the previous section, BAT summary scores were analyzed in which missing data were deleted pairwise, and in which outlying values were "fenced" with extreme valid values. Again, because of the similarities between results for the two datasets, results are presented only for the latter. Consistent with earlier research on the BAT battery, the focus here was on the incremental validity of the BAT beyond prediction of UPT final outcome on the basis of (a) the AFOQT Pilot composite

(PILOT), and (b) previous flying experience (FLYEXP) (Carretta, 1989; Kantor & Bordelon, 1985; Kantor & Carretta, 1988).

Table 11 presents baseline results in which UPT final outcome was regressed simultaneously on PILOT, FLYEXP, and all BAT summary scores using ordinary least squares (OLS) regression. The second column ("Predicted Sign") shows the expected sign of the OLS regression coefficient based on the sign of each predictor's zero-order correlation with UPT final outcome. In six cases, the predictor's regression weight was oppositely signed, indicating the presence of suppressor effects (Lord & Novick, 1968). Also, although the overall regression equation was statistically significant ($F(17,977) = 4.91$, $p < .0001$), many of the individual predictors failed to make a significant incremental contribution to the prediction of UPT final outcome. This was partly due to collinearity among the predictors and the conservative nature of the test for statistical significance of the individual regression weights.

Building on correlational and PCA results presented in the previous two sections, an alternative regression model was estimated which combined many of the predictors shown in Table 11. Specifically, UPT final outcome was regressed simultaneously on:

- PILOT - the USAF's AFOQT Pilot composite;
- FLYEXP - previous flying experience scale;
- AIIR21 - the sum of 21 items in AII having $r \geq .02$ with UPT final outcome;
- ACCUR - Accuracy, the mean of ENCRO, ITMRO, and MRTRO z-scores;
- RT - Response latency, the mean of ENCRT, ITMRT, and MRTRT z-scores;
- TMSSPAC - A cross-product between TMSLD and 1/TMSRT;
- PS2A - Mean of PS2XA and PS2YA z-scores; and
- PS2B - Mean of PS2XB, PS2YB, and PS2ZB z-scores.

Results are shown in Table 12. A greater proportion of these scores contributed incrementally to the prediction of UPT final outcome compared to the baseline regression results shown in Table 11, although not all did (beta weights for RT and ACCUR were zero). The decrease in the R^2 for the revised model compared to the baseline model was small (Change in $R^2 = .01$), which, along with the associated F statistics ($F(17,977) = 4.91$, and $F(8,1536) = 14.13$, for the baseline and revised regression models, respectively) points to the parsimony of the revised regression model.

TABLE 11. BASELINE REGRESSION RESULTS FOR BAT SELECTION BATTERY

Test/ Dimension	Predicted Sign	Beta	p <
PILOT	+	.067	.05
FLYEXP	+	.052	ns
AIIRO	+	-.021	ns
AIIRT	-	.008	ns
ENCRO	+	.025	ns
ENCRT	-	.062	ns
ITMRO	+	-.005	ns
ITMRT	-	-.037	ns
MRTRO	+	-.017	ns
MRTRT	-	-.039	ns
PS2XA	-	-.054	ns
PS2XB	-	.023	ns
PS2YA	-	-.065	ns
PS2YB	-	-.053	ns
PS2ZB	-	-.083	.05
TMSRT	-	-.118	.01
TMSLD	+	.006	ns

Multiple R = .280, $F(17,977) = 4.91$, $p < .01$

Note. Criterion is UPT Final Outcome (1 = Graduated, 0 = Failed to graduate). Cases were deleted pairwise in the event of missing data. Regression coefficients statistical significance were tested by directional hypotheses (i.e., one-tailed t-tests).

TABLE 12. REVISED REGRESSION RESULTS FOR BAT SELECTION BATTERY

Test/ Dimension	Predicted Sign	Beta	$p <$
PILOT	+	.061	.05
FLYEXP	+	.062	.01
AIIR21	+	.048	.05
ACCUR	+	.018	ns
RT	-	-.004	ns
TMSSPLD	+	.108	.01
PS2A	-	-.127	.01
PS2B	-	-.062	.01

Multiple R = .262, F(8,1536) = 14.13, $p < .01$

Note. Criterion is UPT Final Outcome (1 = Graduated, 0 = Failed to graduate). Cases were deleted listwise in the event of missing data. AIIR21 = Mean of 21 items in AII having $r \geq .02$;ACCUR = Mean of ENCRO, ITMRO and MRTRO; RT = Mean of ENCRT, ITMRT, and MRTRT; TMSSPAC = TMSLD x 1/TMSRT; PS2A = Mean of PS2XA and PS2YA; PS2B = Mean of PSYXB, PS2YB, and PS2ZB. Regression coefficients' statistical signifance were tested by directional hypotheses (*i.e.*, one-tailed *t*-tests).

In summary, baseline regression results predicting UPT final outcome from the AFOQT Pilot composite, flying experience, and BAT summary scores (a) evidenced suppressor effects, and (b) showed that few BAT summary scores contributed significant incremental validity in the prediction of UPT final outcome. A revised regression model which combined several BAT summary scores into a smaller number of predictor composites was more parsimonious in the prediction of UPT final outcome, and with little loss in explanatory power compared to the baseline model.

CONCLUSIONS

The purposes of the present research were to investigate:

(a) Internal consistencies of item-level BAT scores;

(b) Alternative test scoring procedures, including methods for the treatment of outlying data points, data transformations, and alternative methods for forming summary scores from item-level data;

(c) The underlying factor structure and differential validity of BAT summary scores;

Results presented in the previous sections suggest the following conclusions:

(a) The internal consistencies of the majority of BAT summary scores are high, indicating that the constructs assessed are measured reliably, and that most tests could be shortened if necessary because of time or budgetary constraints. In some cases (e.g., accuracy scores) lower internal consistencies indicated the unavailability of an appropriate reliability index;

(b) Neither deleting outlying data points nor transforming nonnormal data had a significant impact on either the internal consistency or validity of BAT summary scores in the prediction of UPT final outcome. Nevertheless, it is recommended that outlying summary scores be treated at the summary score level to identify individuals who are not performing according to test instructions;

(c) Very few alternative scoring procedures improved the predictability of UPT final outcome as compared to more routinely calculated summary scores, indicating the usefulness of present methods for forming BAT summary scores;

(d) PCA results indicated that both BAT selection and experimental tests relate meaningfully to lower-order underlying dimensionalities, suggesting some summary scores tap common dimensions of performance; and,

(e) BAT summary scores can be combined into a more efficient model for the prediction of UPT final outcome (in terms of a reduced number of predictors). However, response accuracy and response latency measures failed to make incremental contributions to the prediction of UPT final outcome.

RECOMMENDATIONS

The present study's results reinforce earlier findings of high internal consistency for most BAT scores (e.g., Carretta, 1990a, 1991). The first recommendation is for the maintenance of high levels of reliability in the present and future versions of the test battery. Of course, operational constraints may mandate shorter versions of

certain tests, in which case results in Table 3 can be used to forecast the degree to which measurement reliability would be compromised.

Data transformations, deletion of OIDPs, and alternative methods for forming summary scores from item-level data had little effect overall on increasing summary scores' internal consistencies or correlations with UPT final outcome. These results support the effectiveness of current methods for scoring the BAT battery (Carretta, 1990a) and the idea that statistics developed on normal theory assumptions generally are robust over violation of these assumptions. One general recommendation that emerged from this work is that outliers should be identified at the summary score level using the decision rules outlined in the Appendix in order to identify examinees who appear to not perform according to test instructions. Operationally, outlying scores should be recoded to the score representing the maximum (or minimum) valid score (i.e., outliers should be "fenced" to boundary values defining outlying data).

PCAs and regression results suggest that a smaller number of dimensions than the number of performance scores derived from the BAT is necessary to account for relations among battery scores. This finding is important from the standpoint of achieving a parsimonious model for the prediction of UPT final outcome. The comparison between regression models in Tables 11 and 12 show parsimony in the prediction of UPT final outcome can be achieved without the expense of loss in explanatory power. Thus, the third recommendation is for use of a more parsimonious prediction model such as that shown in Table 12.

REFERENCES

Ackerman, P. L. (1986). Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. *Intelligence, 10,* 101-139.

Adams, J. A. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychological Bulletin, 101,* 41-74.

Belsley, D. A., Kuh, E., & Welsh, R. E. (1980). *Regression diagnostics - identifying influential data and sources of collinearity.* New York: Wiley.

Bentler, P. M., & Woodward, J. A. (1983). The greatest lower bound to reliability. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement.* Hillsdale, NJ: Erlbaum, 237-253.

Bordelon, V. P., & Kantor, J. E. (1986). *Utilization of psychomotor screening for USAF pilot candidates: Independent and integrated selection methodologies.* (AFHRL- TR-86-4). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Campbell, J. P. (1991). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette and L. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd. ed.) (Vol. 1), 687-732.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* Beverly Hills, CA: Sage.

Carretta, T. R. (1987). *Basic Attributes Test (BAT) system: Development of an automated test battery for pilot selection.* (AFHRL-TR-87-9). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Carretta, T. R. (1989). USAF pilot selection and classification systems. *Aviation, Space, and Environmental Medicine, 60,* 46-49.

Carretta, T. R. (1990a). *Basic Attributes Test (BAT): A preliminary comparison between reserve officer training corps (ROTC) and officer training school (OTS) pilot candidates.* (AFHRL-TR-89-50). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Carretta, T. R. (1990b). *Cross-validation of experimental USAF pilot training performance models.* (AFHRL-TR-89-68). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Carretta, T. R. (1991). *Short-term test-retest reliability of an experimental version of the Basic Attributes Test battery.* (AL-TP-1991-0001). Brooks AFB, TX: Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Division.

Carretta, T. R., & Siem, F. M. (1988). *Personality, attitudes, and pilot training performance: Final analysis.* (AFHRL- TP-88-23). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 256-276.

Chatterjee, S., & Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Review, 1,* 379-416.

Chatterjee, S., & Price, B. (1977). *Regression analysis by example.* New York: Wiley.

Cohen, J. & Cohen, C. (1983). *Applied multiple regression/ correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cory, C. H., Rimland, B., & Bryson, R. A. (1977). Using computerized tests to measure new dimensions of abilities: An exploratory study. *Applied Psychological Measurement, 1,* 101-110.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart, & Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Dolgin, D. L., & Gibb, G. D. (1989). Personality assessment in aviator selection. In R. S. Jensen (Ed.), *Aviation psychology.* Aldershot: Gower, 288-320.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan, 105-146.

Games, P. A. (1983). Curvilinear transformations of the dependent variable. *Psychological Bulletin, 93,* 382-387.

Games, P. A. (1984). Data transformations, power, and skew: A rebuttal to Levine and Dunlap. *Psychological Bulletin, 95,* 345-347.

Greaud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10,* 23-34.

Green, B. F. (1988). Construct validity of computer-based tests. In H. Wainer & H. I. Braun (Eds.), *Test validity.* Hillsdale, NJ: Erlbaum, 77-86.

Hattie, J. (1985). Methodology review: Assessing the unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Kantor, J. E., & Bordelon, V. P. (1985). The USAF pilot selection and classification research program. *Aviation, Space, and Environmental Medicine, 56,* 258-261.

Kantor, J. E., & Carretta, T. R. (1988). Aircrew selection systems. *Aviation, Space, and Environmental Medicine, 59*(11, Supplement), A32-A38.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Levine, D. W., & Dunlap, W. P. (1983). Data transformation, power, and skew: A rejoinder to Games. *Psychological Bulletin, 93,* 596-599.

Link, S. W. (1982). Correcting response measures for guessing and partial information. *Psychological Bulletin, 92,* 469-486.

Long, J. S. (1983). *Confirmatory factor analysis.* Beverly Hills, CA: Sage.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Luce, R. D. (1986). *Reaction times: Their roles in inferring elementary mental organization.* New York: Oxford.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43,* 335-354.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression.* Reading, MA: Addison-Wesley.

Mulaik, S. A. (1972). *The foundations of factor analysis.* New York: McGraw-Hill.

Mulaik, S. A. (1988). Confirmatory factor analysis. In J. R. Nesselroade and R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed.). New York: Plenum, 259-288.

Novello, J. R., & Youssef, Z. I. (1974). Psycho-social studies in general aviation: I. Personality profile of male pilots. *Aerospace Medicine, 45*(2) 185-188.

Nunnally, J. C. (1970). *Introduction to psychological measurement.* New York: McGraw-Hill.

Ollman, R. (1966). Fast guesses in choice reaction time. *Psychonomic Science, 6,* 155-156.

Pachella, R. G. (1974). The interpretation of reaction time in information-processing research. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition.* Hillsdale, NJ: Erlbaum, 41-82.

Passey, G. E., & McLaurin, W. A. (1966). *Perceptual-psychomotor tests in aircrew selection: Historical review and advanced concepts.* (PRL-TR-66-4). Lackland AFB, TX: Personnel Research Laboratory.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research.* Ft. Worth, TX: Holt, Rinehart, and Winston.

Ree, M. J. (1976). *Effects of item option weighting on the reliability and validity of the AFOQT for pilot selection.* (AFHRL-TR-76-76). Lackland AFB, TX: Air Force Human Resources Laboratory, Personnel Research Division.

Sanders,J. H. Jr., Valentine, L. D. Jr., & McGrevy, D. F. (1971). *The development of equipment for psychomotor assessment.* (AFHRL-TR-71-40). Lackland AFB, TX: Personnel Research Laboratory.

Siem, F. M. (1990). *Predictive validity of an automated personality inventory for Air Force pilot selection.* (AFHRL-TP-90-55). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Siem, F. M., Carretta, T. R., & Mercatante, T. A. (1988). *Personality, attitudes, and pilot training performance: Preliminary analyses.* (AFHRL-TP-87-62) Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Skinner, H. A. (1978). The art of exploring predictor-criterion relationships. *Psychological Bulletin, 85,* 327-337.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95,* 334-344.

Stevens, J. P. (1986). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Erlbaum.

Wainer, H., & Braun, H. I. (1988). *Test validity.* Hillsdale, NJ: Erlbaum.

Wickens, C. D. (1984). *Engineering psychology and human performance.* Glenview, IL: Scott, Foresman, 342-346 and 362-364.

Wittman, W. W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.) *Handbook of multivariate experimental psychology* (2nd ed). New York: Plenum, 505-560.

Yellott, J. I. Jr. (1967). Correction for guessing in choice reaction time. *Psychonomic Science, 8,* 321-322.

Yellott, J. I. Jr. (1971). Correction for fast guessing and the speed-accuracy tradeoff in choice reaction time. *Journal of Mathematical Psychology, 8,* 159-199.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research, 17,* 253-269.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432-442.

APPENDIX A: DECISION RULES FOR TREATMENT OF OUTLIERS

TABLE A-1. DECISION RULES FOR TREATMENT OF OUTLIERS

Test	Decision Rule
AIA	Item Level:
	AIART1 TO AIART81 - LOWEST THRU 200 = MISSING
	AIART1 TO AIART81 - 12800 THRU HIGHEST = MISSING
	Summary Score Level:
	AIARO - LOWEST THRU .30 = .30
	AIART - LOWEST THRU 1549 = 1549
ENC	Item Level:
	NCB1RT2 TO NCB1RT64, NCB2RT1 TO NCB2RT63, and NCB3RT1 TO NCB3RT63 - LOWEST THRU 5.29 = MISSING
	NCB1RT2 TO NCB1RT64 - 7.28 THRU HIGHEST = MISSING
	NCB2RT1 TO NCB2RT63 - 7.31 THRU HIGHEST = MISSING
	NCB3RT1 TO NCB3RT63 - 7.63 THRU HIGHEST = MISSING
	Summary Score Level:
	ENCRO - LOWEST THRU .75 = .75
	ENCRT - 1500 THRU HIGHEST = 1500
ITM	Item Level:
	ITB2RT1 TO ITB6RT24 - 7.78 THRU HIGHEST = MISSING
	ITB2RT1 TO ITB6RT24 - LOWEST THRU 5.29 = MISSING
	Summary Score Level:
	ITMRO - LOWEST THRU .75 = .75
	ITMRT - 2000 THRU HIGHEST = 2000
MRT	Item Level:
	MRB1RT1 TO MRB3RT48 - LOWEST THRU 200 = MISSIMG
	MRB1RT1 TO MRB1RT24 - 3233 THRU HIGHEST = MISSING
	MRB1RT25 TO MRB1RT48 - 3361 THRU HIGHEST = MISSING
	MRB2RT1 TO MRB2RT24 - 3220 THRU HIGHEST = MISSING
	MRB2RT25 TO MRB2RT48 - 3386 THRU HIGHEST = MISSING
	MRB3RT1 TO MRB3RT24 - 2856 THRU HIGHEST = MISSING
	MRB3RT25 TO MRB3RT48 - 3127 THRU HIGHEST = MISSING

Summary Score Level:

MRTRO - LOWEST THRU .50 = .50
MRTRT - 2500 THRU HIGHEST = 2500

PS2 Item Level:

PS2XA1 TO PS2XA10 - 9929 THRU HIGHEST = MISSING
PS2YA1 TO PS2YA10 - 17750 THRU HIGHEST = MISSING
PS2XB1 TO PS2XB10 - 13500 THRU HIGHEST = MISSING
PS2YB1 TO PS2YB10 - 12125 THRU HIGHEST = MISSING
PS2ZB1 TO PS2ZB10 - 10800 THRU HIGHEST = MISSING

Summary Score Level:

PS2XA - 9000 THRU HIGHEST = 9000
PS2YA - 10000 THRU HIGHEST = 10000
PS2XB - 12500 THRU HIGHEST = 12500
PS2YB - 10000 THRU HIGHEST = 10000
PS2ZB - 9000 THRU HIGHEST = 9000

TMS Item Level:

RTT11I1 TO RTT13I6, RTT14I1 TO RTT14I5,
RTT15I1 TO RTT16I6 - LOWEST THRU 200 = MISSING

RTT11I1 TO RTT13I6, RTT14I1 TO RTT14I5,
RTT15I1 TO RTT16I6 - 3000 THRU HIGHEST = MISSING

Summary Score Level:

TMSRT - 2000 THRU HIGHEST = 2000
TMSLD - LOWEST THRU 75 = 75

ABC Item Level:

ABCST1 TO ABCST48 - LOWEST THRU 200 = MISSING
ABCRT1 TO ABCRT48 - LOWEST THRU 200 = MISSING
ABCRT1 TO ABCRT48 - 7900 THRU HIGHEST = MISSING

Summary Score Level:

ABCRO - LOWEST THRU .15 = .15
ABCST - 30000 THRU HIGHEST = 30000
ABCRT - 4000 THRU HIGHEST = 4000
ABCCO - LOWEST THRU 3 = 3

ANT Item Level:

ANT1 TO ANT10 - LOWEST THRU -92 = MISSING
ANT11 TO ANT20 - LOWEST THRU -100 = MISSING
ANT21 TO ANT30 - LOWEST THRU -133 = MISSING
ANT31 TO ANT40 - LOWEST THRU -100 = MISSING
ANT41 TO ANT50 - LOWEST THRU -150 = MISSING

Summary Score Level:

ANTTE - LOWEST THRU -75 = -75

PAT Item Level:

PATRT1 TO PATRT30 - LOWEST THRU 200 = MISSING
PATRT1 TO PATRT30 - 10600 THRU HIGHEST = MISSING

Summary Score Level:

PATRO - LOWEST THRU .40 = .40
PATRT - LOWEST THRU 1500 = 1500

SAA Item Level:

SI1T1I1 TO SI3T4I15 - 10 THRU HIGHEST = MISSING
CI1T1I1 TO FI3T4I15 - 3600 THRU HIGHEST = MISSING

Summary Score Level:

SAAER - 3600 THRU HIGHEST = 3600
